A mixed-effect state space model to environmental data

Marco Costa and Magda Monteiro

Citation: AIP Conference Proceedings **1648**, 110002 (2015); doi: 10.1063/1.4912409
View online: http://dx.doi.org/10.1063/1.4912409
View Table of Contents: http://scitation.aip.org/content/aip/proceeding/aipcp/1648?ver=pdfcov
Published by the AIP Publishing

A Mixed-effect State Space Model to Environmental Data

Marco Costa[*,†] and Magda Monteiro[*,**]

[*]*Escola Superior de Tecnologia e Gestão de Águeda, Universidade de Aveiro, 3754–909 Águeda, Portugal*
[†]*Centro de Matemática e Aplicações Fundamentais, Universidade de Lisboa, 1649-003 Lisboa, Portugal*
[**]*Centro de Investigação e Desenvolvimento em Matemática e Aplicações, Universidade de Aveiro, 3810-193 Aveiro, Portugal*

Abstract. This work presents some common issues in the statistical analysis of time series of environmental area. The discussion and the presentation of solutions is raised by the study of a time series of the oxygen concentration variable in a water quality monitoring site in the river Vouga hydrological basin in Portugal. Issues such as trends, seasonality, temporal correlation and detection of change points are addressed.

Keywords: Time series analysis, Kalman filter, state space model, environmental data.
PACS: 02.50.Ey, 05.45.Tp

INTRODUCTION AND DESCRIPTION OF THE DATA

The time series analysis of environmental data has been an important tool on the assessment and monitoring of natural processes. However, the statistical analysis of environmental data raises a number of issues related to the characteristics of this type of data. Oftentimes, the hydrometeorological and environmental data are monthly observations with a strong seasonal character. Some analyzes can circumvent this problem considering the annual averages, if it is appropriate ([9]). The Gaussian assumption for the data distribution is not suitable in many applications. This question is relevant due to many statistical procedures assume this assumption. When data distribution are skewed, the observations can be transformed, usually, using the logarithmic function. However, this procedure has problems in the interpretation of the models and their parameters. A temporal correlation structure on environmental data are very common. For instance, [6] applies state space models in order to accommodate the time dependence of monthly water quality variables, [1] presents inference propers of regression models which residuals follow any autoregressive stationary process. One of the most common properties is the non-stationarity of a time series. The non-stationarity can assume many forms. For instance, the existence of a trend or heterogeneity in variance, the change of a property in a known instant or in a unknown point (usually designed as a change-point). An environmental time series may have some of these properties simultaneously implying that its analysis should incorporate various techniques in an appropriate way. This work presents the modeling of a water quality variable time series which presents a set of properties. This example allows to expose statistics issues on the time series modeling and it presents some solutions for them. This paper examines the monthly data of the Dissolved Oxygen (mg/l) in the Carvoeiro water monitoring site in the hydrological basin of the river Vouga, Portugal. Data were collected from the SNIRH (Sistema Nacional de Recursos Hídricos), the national information system for water resources of Portugal, at `http://snirh.apambiente.pt/`. The modeling of water quality variables will be performed using both linear regression ([5, 11]) and state space models ([3, 6]). Linear models are usually applied for their simplicity and well known statistical properties. State space models associated with Kalman filter allow incorporating some components in the model in a dynamic point of view and the Kalman filter produces forecasts with the respective confidence intervals.

REGRESSION MODELING

The exploratory analysis of data shows that there are a clear structural change in the trend during in the year of 2000 (see Figure 2). Instead of considering two separate analysis, we intend to find a global model that accommodates this property. Thus, it was followed the methodology adopted in [2]. Considering the DO variable as $Y_t, t = 1, ..., 283$, after fitting a regression model with a quadratic trend up to a certain time t_0 to all possible month/year and a constant to the

Proceedings of the International Conference on Numerical Analysis and Applied Mathematics 2014 (ICNAAM-2014)
AIP Conf. Proc. 1648, 110002-1–110002-4; doi: 10.1063/1.4912409

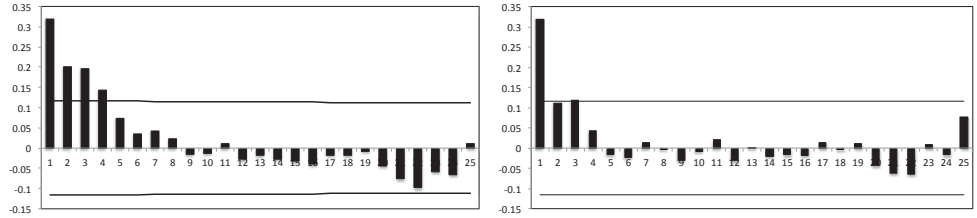

FIGURE 1. Sample autocorrelation function and sample partial autocorrelation function of the residuals of the regression analysis.

same series from the next month/year, we choose the one with a smaller residual sum of squares. This procedure led to $t_0 = 139$, i.e., October 2000 and the estimated model of the trend was

$$T_t = \begin{cases} -0.0014t^2 + 0.01986t + 6.8991 & if \quad t \leq 139 \\ 9.128 & if \quad t > 139. \end{cases} \tag{1}$$

From residuals the seasonal coefficients were estimated as suggested by [9], i.e., the easiest way to handle the seasonality is to subtract from each month the overall average of this month,s_{r+12k}, with $r = 1, ..., 12$ and $k = 0, 1, ...$ As will be seen this procedure is sufficient for this stage of modeling. On the one hand, from Figure 1, which represents both the sample autocorrelation function and the partial autocorrelation function, it is clear that the residuals series follows an AR(1) process, that is, there is a temporal correlation structure. On the other hand, we concluded that subseries variances of each month of the year are heterogeneous and they tend to have greater variability associated to the highest seasonal coefficients (sample linear correlation coefficient equal to 0.38).

A MIXED-EFFECT STATE-SPACE MODEL

In order to incorporate the temporal correlation and to accommodate the heterogeneity of the variances of the sub series of each month, we adopte a mixed-effect state-space model. The model is defined as

$$Y_t = (T_t + S_t)\beta_t + e_t \tag{2}$$

$$\beta_t = \mu + \phi(\beta_{t-1} - \mu) + \varepsilon_t \tag{3}$$

where the trend component T_t has a structure defined in (1), S_t is a periodic function with the twelve averages as above, the process $\{\beta_t\}$ is a stationary AR(1) with Gaussian errors ε_t with mean μ and e_t is the observation equation error, assumed to be a Gaussian white noise process. The model has two main components: a regression structure which incorporate both trend and seasonality and an unobservable process $\{\beta_t\}$, the *state*. It is assumed that state process will calibrate the deterministic structure $T_t + S_t$. Another important feature of the model is that by its own formulation it allows the variances heterogeneity previously identified. As the state process $\{\beta_t\}$ is unknown, it must be predicted by the Kalman filter recursions. The Kalman filter provides optimal unbiased linear one-step-ahead predictions of the unknown states and their corresponding mean square error (MSE), which are denoted by $\widehat{\beta}_{t|t-1}$ and $P_{t|t-1}$ respectively. Furthermore, the Kalman filter provides optimal filtered predictions $\widehat{\beta}_{t|t}$ and their MSE $P_{t|t}$. If the errors are further assumed to have a conditional Normal distribution, then $\widehat{\beta}_{t|t-1}$ is the conditional mean of β_t. Model's parameters $\Theta = (\mu, \phi, \sigma_e^2, \sigma_\varepsilon^2)$ are obtained through the maximum likelihood estimation ([7]). Table 1 presents the maximum likelihood estimates of Θ. In this paper we adopted a decomposition approach type to estimate the model's parameters, that is, estimating parameters step-by-step in each phase of the modeling procedure. However, it is possible to estimate the model (2)-(3) through the EM-algorithm method (see for instance [10]). The forecasts one step-ahead with the respective empirical confidence interval at 95% were obtained by Kalman filter equation, that is, $\widehat{Y}_{t|t-1} \pm 1.96\sqrt{MSE_{\widehat{Y}_{t|t-1}}}$. Figure 2 shows the observed DO concentration, the forecast one step-ahead with the respective confidence interval at 95%. The percentage of observations outside of the respective empirical confidence interval is 6.36% (18/283). The residuals analysis shows that there are some outliers; the biggest residual is respected

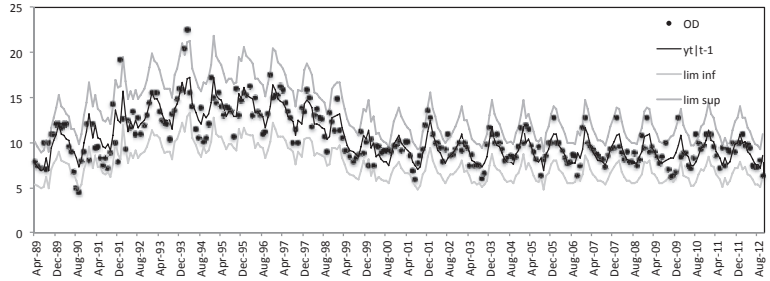

FIGURE 2. Dissolved oxygen concentration and the forecast one step-ahead with the respective confidence intervals at 95%.

TABLE 1. Gaussian maximum likelihood estimates of the state space model.

$\widehat{\mu}$	$\widehat{\phi}$	$\widehat{\sigma}_{\varepsilon}^2$	$\widehat{\sigma}_{e}^2$
1.00077	0.43001	0.01790	0.24292

to January 1992. The p-value of the Kolmogorov-Smirnov test to the normality is equal to 2.8% leading to rejection of null hypothesis. However, the Gaussian distribution is not rejected (K-S p-value=20.0%) when the largest outliers are not considered.

CHANGE POINT DETECTION

The modeling procedure exposed before shows a change in the trend component of the DO concentration. This fact is clear from the exploratory analysis. However, other structural changes are not so evident and they need a statistical analysis in order to evaluate their existence. The model (2)-(3) established a calibration procedure considering $T_t + S_t$ a intrinsic component of the variable which, in a dynamic way, is calibrate by the state process $\{\beta_t\}$. Thus, it is reasonable to investigate changes in this structure which may be relevant in the water monitoring process. For this purpose, the filtered predictions of the state process, $\widehat{\beta}_{t|t}$ can be analyzed relatively to the existence of a change (or more) in their mean as the best prediction of an AR(1) process. Since the sample size is considerable, the results on the Gaussian distribution discussed before was devalued. We used test statistics that are called *maximum type* and referred in [4]. The null hypothesis claims that X_1, X_2, ..., X_n are distributed according to $N(\mu, \sigma^2)$. The alternative claims that there exists a time point $k \in \{1, 2, ..., n-1\}$ such that X_1, X_2, ..., X_k are distributed according to $N(\mu_1, \sigma^2)$ and X_{k+1}, ..., X_n are distributed according to $N(\mu_2, \sigma^2)$ with $\mu_1 \neq \mu_2$. Supposing σ^2 is unknown, the problem described can be treated based on two-sample t tests. The test statistic $T(n)$ is the maximum of the absolute values of two sample t-test statistics

$$T(n) = \max_{1 \leq k < n} |T_k| = \max_{1 \leq k < n} \sqrt{\frac{(n-k)k}{n}} \left| \overline{\beta}_{k|k} - \overline{\beta}^*_{k|k} \right| \frac{1}{s_k}$$

where $\overline{\widehat{\beta}}_{k|k} = \frac{1}{k}\sum_{i=1}^{k} \widehat{\beta}_{i|i}$, $\overline{\widehat{\beta}}^*_{k|k} = \frac{1}{n-k}\sum_{i=k+1}^{n} \widehat{\beta}_{i|i}$ and $s_k = \sqrt{\frac{1}{n-2}\left[\sum_{i=1}^{k}\left(\widehat{\beta}_{i|i} - \overline{\widehat{\beta}}_{k|k}\right)^2 + \sum_{i=k+1}^{n}\left(\widehat{\beta}_{i|i} - \overline{\widehat{\beta}}^*_{k|k}\right)^2\right]}$.

The test statistic $T(n)$ were computed considering the sequences of $\beta_{t|t}$. Based on [9], the interpolated value to $n = 283$ is 3.207. However, this critical value assumes that the observations are uncorrelated. So, it is necessary to correct this critical value according to [8]. Thus, in order to obtain 5% critical values, the value 3.207 is multiplied by $[(1+\widehat{\phi})(1-\widehat{\phi})^{-1}]^{1/2}$, where $\widehat{\phi}$ is obtained in Table 1. Thus, 5% critical value is 5.079. Table 2 presents the results from a binary segmentation procedure (proposed initially by [12]) in order to detect multiple change point. This procedure ends when the observed value of the test statistics is less than the respective critical value (the critical value is obtained by interpolation according to the subsample size). This procedure identified two change points, April 1990 and December 1991. Figure 3 represents the change points and the three mean levels of the calibration factors process. For chronological order, the mean levels of the process $\{\beta_t\}$ are 1.190, 0.835 and 1.005. This means that in the first

TABLE 2. Results from the binary segmentation procedure.

n_i	$T(n_i)_{obs}$	month/year	**critical value not corrected**	**critical value corrected**
283	5.1893	Apr-90	3.2066	5.0790
270	5.7600	Dec-91	3.2040	5.0749

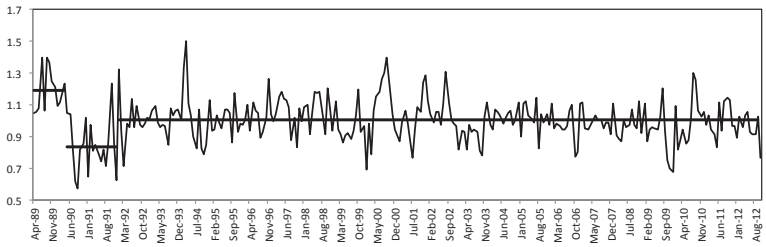

FIGURE 3. Representation of the change points and the three mean levels of the $\{\widehat{\beta}_t\}$.

period of time the DO variable was, on average, 19% higher than those predicted by the regression component. In the second period of time the observations were, on average, 16.5% lower and in last period of time the observation were, on average, very close to the regression model predictions. From the environmental point of view, the analysis of the calibration factor in each time can be relevant in order to identify unanticipated changes.

ACKNOWLEDGMENTS

Marco Costa was partially supported by Fundação para a Ciência e a Tecnologia, PEst OE/ MAT/ UI0209/ 2011. Magda Monteiro was partially supported by Portuguese funds through the CIDMA - Center for Research and Development in Mathematics and Applications, and the Portuguese Foundation for Science and Technology ("FCT–Fundação para a Ciência e a Tecnologia"), within project PEst-OE/MAT/UI4106/2014.

REFERENCES

1. T. Alpuim, and A. El-Shaarawi, On the efficiency of regression analysis with AR(p) errors, *Journal of Applied Statistics* **35**, 717–737 (2008).
2. T. Alpuim, and A. El-Shaarawi, Modeling monthly temperature data in Lisbon and Prague, *Environmetrics* **20**, 835–852 (2009).
3. M. Costa, and A.M. Gonçalves, Clustering and forecasting of dissolved oxygen concentration on a river basin, *Stoch. Env. Res. Risk A.* **25**, 151–163 (2011).
4. M. Costa, and A.M. Gonçalves, Application of Change-Point Detection to a Structural Component of Water Quality Variables, In *Numerical Analysis and Applied Mathematics ICNAAM 2011*, edited by S. Theodore et al., AIP Conference Proceedings 1389, American Institute of Physics, New York, 2011, pp. 1565–1568.
5. A.M.. Gonçalves, and T. Alpuim, Water Quality Monitoring using Cluster Analysis and Linear Models, *Environmetrics* **22**, 933–945 (2011).
6. A.M. Gonçalves, and M. Costa, Predicting seasonal and hydro-meteorological impact in environmental variables modelling via Kalman filtering, *Stoch. Env. Res. Risk A.* **27**, 1021–1038 (2013).
7. A. Harvey, *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press, Cambridge.
8. D. Jarušková, Change-point detection meteorological measurement, *Mon Weather Rev* **124**, 1535–1543 (1996).
9. D. Jarušková, Some problems with application of change-point detection methods to environmental data, *Environmetrics* **8**, 469–483 (1997).
10. P. Kokic, S. Crimp, and M. Howden, Forecasting climate variables using a mixed-effect state-space model, *Environmetrics* **22**, 409–419 (2011).
11. J.A. Renwick, A.B. Mullan, and A. Porteous, Statistical downscaling of New Zealand climate, *Weather and Climate* **29**, 24–44 (2009).
12. L.J. Vostrikova, Detecting "disorder" in multidimensional random processes, *Soviet Mathematics Doklady* **24**, 55–59 (2009).